Information Architecture Sponsors Data Warehousing Study

As part of its effort to improve the sharing of information and the effective use of emerging technologies, the Information Architecture (IA) project is sponsoring a study of data warehousing. Data warehousing is a method of making historical data more easily accessible to the Laboratory-wide community. As its name indicates, a data warehouse is a repository for storing information. Figure 1 illustrates how the data warehouse takes information from a variety of sources and stores it for access by a variety of users.

The purpose of a data warehouse is to assist decision makers by making information easily accessible in a useful way. The data warehouse is a decision-making tool, rather than a day-to-day operations tool.

The concept of data warehousing has recently been successfully implemented by a number of businesses to assist with decision making problems faced by managers. A special challenge for the Laboratory IA teams is to extend this corporate data warehouse model for use by all Laboratory employees, including scientists, engineers, and administrators. Several key elements concerning a data warehouse are described below.

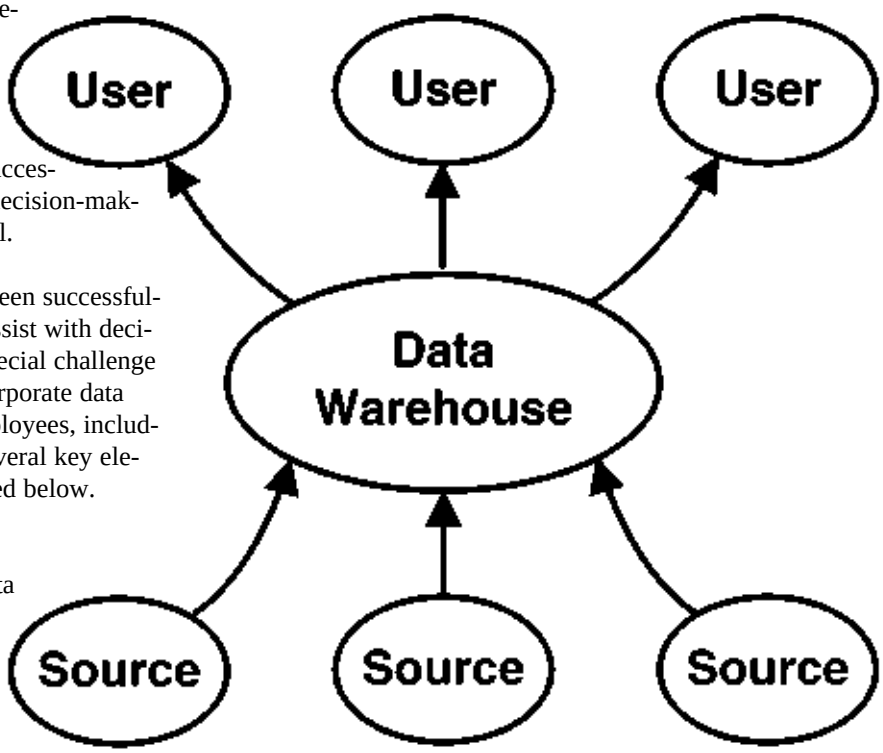

Figure 1. Data Warehouse Information Flow

One-Way Only Information Flow

The information flows from its source to the data warehouse, but the data warehouse never updates its sources. Similarly, users can access the information in the data warehouse, but they cannot update it. Consequences of the one-way flow include the following:

- Users can manipulate data at their desktop without affecting the source data, providing enhanced assurance of source data integrity.
- Data can be optimized for query and otherwise transformed for ease and speed of use.
- Data must include a time element in order to remain meaningful for future use.

The Independent Data Warehouse

The data warehouse is independent of both the sources of information and the user tools that access it. This allows the data warehouse to easily accommodate varying source and access tools. More specifically,

- The data warehouse can accept data from a variety of sources in a variety of formats, including transaction systems that feed alphanumeric data, optical systems that feed imaging data, and more.

- The data warehouse can provide data to a wide range of client platforms, including PCs, Macintoshes, UNIX workstations, and more.

- The data warehouse can adapt easily to both existing and emerging source and client technologies.

- By integrating data from many sources and storing it historically, the data warehouse provides a foundation for fact-based trend analysis.

Flexible Data Formats

As envisioned by the IA data warehouse standards development team, there is no reason to restrict the types of data that can be stored in the warehouse. Because the warehouse stands independent from its sources and clients, it can accommodate new formats without affecting existing systems. Some of the many types of data that have been identified by the team are

- Standard alphanumeric data

- Images in varying formats

- Sound and multimedia

- PR and training movies

- Geographical images

- Hypertext-capable documents

As new ways of storing data emerge, they too can be accommodated. This flexibility will help the data warehouse take full advantage of both existing and emerging technologies.

Data Transformation and Optimization

The data stored in the warehouse will rarely be identical in form to the data in its source. If, for example, the source of the data is a real-time transaction system, then the requirements of the real-time system will be different from the requirements of the data warehouse.

Real-time transaction or "operations" systems contain the current state of information, and are designed for rapid, accurate updates with multiple assurances of data integrity. By contrast, a data warehouse allows easy ad hoc access to information with a time content, providing meaningful information long after the current state of the data has changed.

For example, a data warehouse might be used to track funding proposals made by Laboratory employees. Information collected from other systems could include types of proposals made, proposals accepted, skills needed to perform the work, resulting publications, etc. The data warehouse would then provide a picture (showing change over time) of the types of proposals that are successful, changes in needed skills at the Laboratory, and the types and number of publications that we produce. This kind of information should be useful not only to senior management, but to all employees involved in preparation of proposals.

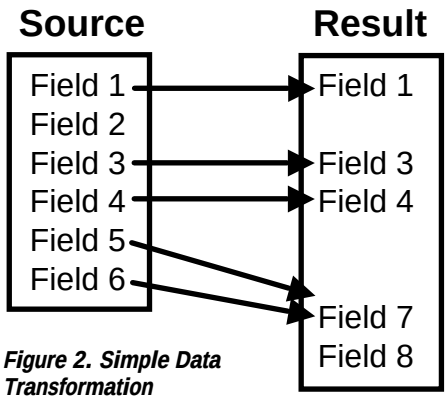

Figure 2. Simple Data Transformation

Figure 2 depicts a simple model of transforming data from a real-time transaction system to the data warehouse. Some information is deleted, some is combined, and some is added.

The transformation can also occur in different ways. A document with illustrations, for example, might be translated from the particular format it was created under to a read-only, compressed format such as Adobe's portable document format (PDF).

Whatever the particular changes, the transformation allows the historical data to be tuned for ease of ad hoc access, speed of access, consistency, and meaningfulness.

Process Overview

The data warehouse team will begin by constructing a logical model, as shown in Figure 3, which will lead to the physical model.

The logical model is a theoretical description of what the data warehouse should do, while the physical model is a practical description of how the data warehouse will be implemented.

As a theoretical description, the logical model can focus on an ideal state, unencumbered by the limitations of current-state technology or financing. It begins as a more detailed expression of the IA principles that were adopted by the Laboratory Leadership Council in May, especially the following:

• Shared information is the foundation of a unified Laboratory.

• Access to information is the rule, not the exception.

• Laboratory information and data are corporate assets and are managed accordingly.

In cooperation with the IA metrics team, the data warehouse team will analyze specific customer needs from the Laboratory-wide community. These, when combined with surveys of how external laboratories and companies are using data warehouses, will form the basis of the logical model.

After the team has completed its description of the ideal state, it will turn its attention to how the logical model can actually be implemented. This work will include analysis of the current state of both technology and the Laboratory, which will serve as a reality-check when constructing the physical model.

The data warehouse team anticipates that the physical model will initially be expressed through a pilot project. This project will be used to validate the practical benefits of the data warehouse. Additional news about the pilot project can be anticipated at a later date.

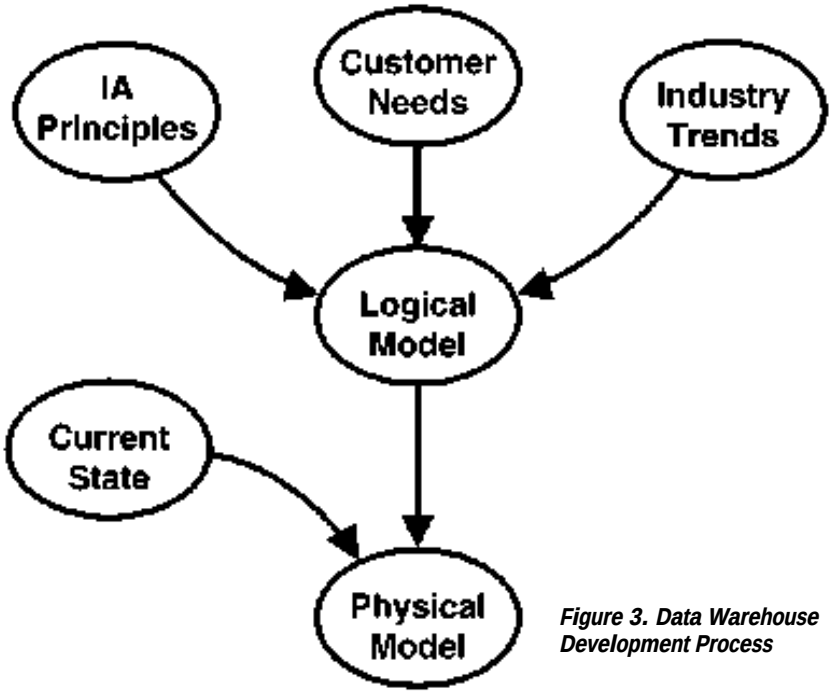

Figure 3. Data Warehouse Development Process

Data Warehouse Guiding Principles (preliminary)

The IA data warehouse team has developed a preliminary set of ten high-level principles to guide the design of a data warehouse for the Laboratory. These principles, listed below, will undergo continuing revisions as work on the project continues.

1. The Laboratory data warehouse is based on a sound corporate data model.

2. Technical systems and services are managed separately from information content.

3. Stewardship of data is uniquely assigned to those organizations that are the business owners of the information.

4. The data warehouse is a single logical application residing on multiple servers and security partitions.

5. The data warehouse is tied into LCID, the Laboratory Corporate Information Directory.

6. The data warehouse is designed to be flexible, taking advantage of such technologies as multi-platform, multi-tier, distributed architecture. As such, it adapts to both evolving business rules and emerging technologies.

7. Usage and performance measures are designed into the data warehouse.

8. Security of the data warehouse is based upon access control and data, not application level security.

9. Encryption and digital signatures may be future enabling technologies for the data warehouse.

10. Data is uploaded into the data warehouse from many sources in many formats and in varying degrees of automation.

IA Team Leaders Stress Importance of Laboratory-Wide Input

A lunch time conversation with two of the team leaders from the IA effort shows that both feel the promise of the IA project lies in its attention to Laboratory-wide customer input and consensus building.

"In general," says Steve Blair (NIS-3) leader of the IA data warehouse team, "the role of the information architecture is to identify problems in how we use information tools, recognize that we can fix the problems, and then get the right people together so we can solve them."

Karl-Heinz Winkler, DDCIC, co-leader of the overall IA project, agrees. He adds, "The institution is looking to us to come up with a plan, recognizing the realities of life at the Laboratory."

Both point to the diverse membership of the IA data warehouse team, which currently includes representatives from eight Laboratory divisions.

Winkler continues, "We are an empowered, self-directed team that makes decisions through consensus. We don't make a decision unless there is really an issue, and we work together to reach resolution in a mutually acceptable, mutually beneficial way."

Blair stresses that the team composition and operating principles free the team to effectively approach problems.

He says, "Because we are not 'owned' by any single division, we can invite people to step beyond their parochial concerns. We can elevate the discussion, get people to apply the same openness of mind they use in scientific problems, challenge people, and make it fun again."

The novelty of data warehousing offers further benefits to the process. "Because it is new," Winkler says, "you don't already have set ways of doing things. You aren't encumbered by current practices, and you are better able to get the right people together to develop the organizing structure to address the systemic issues."

"This leads to two key points," Blair adds. "First, we must ask whether we are tackling important problems for the Laboratory and whether we are applying the right technology. In the case of the data warehouse, we are addressing significant problems and there is as of yet no better alternative approach.

"And second," he concludes, "we need to extend the data warehouse model and make it useful to scientists and engineers—in addition to administrators—because the Laboratory's business is in fact science and technology."

For further information about the IA data warehouse project, look on-line via Gopher/Mosaic under "Computing at LANL/ Information Architecture Project," or contact Steve Blair (NIS-3) 665-5895 or sgblair@lanl.gov.

Tad Lane, tad@lanl.gov, (505) 667-0886
Communications Arts and Services (CIC-1)